지능형 SoC와 그 응용

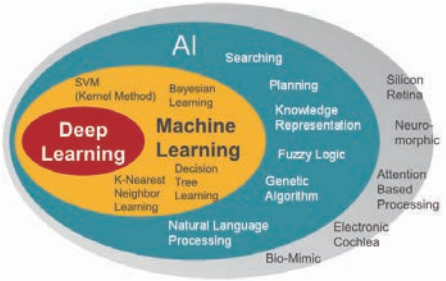
Ⅰ. 서 론

현대를 정보혁명시대라 부른다. 현대에는 우리가 눈치 채고 있든 모르고 있든, 물건 즉 원자, 분자로 된 제품만이 아니라 비트로 된 디지털 정보들이 상품이 되어 부가가치를 창출하고 있다. 무게와 형체가 있는 물건들, 즉 Atom을 가공, 포장하여 팔고 사는 경제 시대로부터 정보의 생산, 전송, 가공, 표시와 소비의 각 부분과 그 흐름이 경제의 주된 핵심으로 자리 잡고 있다. 이 정보에는 데이터와 그 데이터의 1차 가공품인 "정보(Information)" 그리고 정보를 재가공한 "지식"과 지식을 바탕으로 한 "지능"이라는 Hierarchy가 존재한다.

이제까지의 정보혁명의 원동력은 컴퓨터와 통신을 중심으로 한 Data와 Information의 자동화였다. 현존하는 대부분의 전자기기들이 Data와 Information을 처리하는 기계들이다. 정보혁명에서 취급하는 정보는 단순히 모아 놓은 데이터들로 여러분들의 컴퓨터 파일들을 보면 알 수 있듯 시간이 지나면서 점점 더 늘어가는 데이터들과 불규칙한 파일들에 의해 혼란스러워지며 웹에서 떠도는 정보들과 이를 처리하는 기기들은 정보들이 늘어 가면 갈수록 사용자들은 무엇을 선택하여야 할지 더욱 혼돈에 빠져 "디지털 미아"가 될 수밖에 없다. 마치 감각기관이 발달한 생명체가 많은 정보들을 외부로부터 받고 있지만 "뇌"와 같은 지능기관이 없어 혼란스러워 하고 있는 것과 같다.

이제 바야흐로 지능혁명 시대가 도래했다. Data와 Information을 다루던 시대를 지나 Knowledge와 Intelligence를 다룰 수 있는 스마트 머신들이 주가 된다. 스마트 기기들은 사용자의 취향을 학습하고 그들의 필요를 예측하며 어느 정도 자율적으로 행동하는 지능형 시스템이 되어야 한다. 이 때에는 센싱(Sensing)기능과 통신 기능이 이전의 컴퓨터와 통신기기만이 아니라 사람이나 자동차 및 비행체 그리고 로봇 등

유회준
KAIST 전기 및 전자공학부



〈그림 1〉 AI와 Machine Learning, Deep Learning의 관계 및 관련 연구

의 모든 물체(IoT)들에 내재화되어 있으며 지능화 기능이 추가 되어 스마트 머신이 되는 것이 특징이다. 지능혁명의 한 가지 예가 바로 자동차 내비게이터이다. 예전에는 지도를 보고 사람들에게 물어서 가야 했던 길들을 내비게이터의 지시에 의해 단순히 운전만을 하면 목적지에 도착할 수가 있게 되었다. 이 경우 예전의 지능적인 행동, 즉 목적지까지의 최단 거리 또는 각종 교통신호와 도로 조건들에 대한 고려는 내비게이터가 맡아서 처리하고 사람은 단순한 동작만을 수행하면 임무가 완수된다. 이렇게 지능적인 일들을 기계의 도움을 받아서 처리하게 되는 것이 바로 지능혁명인 것이다. 이러한 분야는 앞으로 요리안내, 건강관리, 학습관리 및 여행지 안내 등 의식주 및 우리의 생활 전반에 널리 보급될 것이다. 기존의 정보를 사용자의 현재의 필요에 맞는 지능으로 바꾸어 장소와 시간에 구애됨이 없이 지능 서비스를 받을 수 있도록 하는 제품들이 다양한 모습을 지닌 채 우리를 찾아 올 것이다.

이러한 지능을 연구하는 분야로는 〈그림 1〉에서 보듯이 AI 또는 Machine Learning이 알려져 있으며 최근 들어서는 Deep Neural Network 또는 Deep Learning이 새로운 바람을 일으키며 다양한 영역에서 지능혁명을 일으키고 있다. 예를 들자면 컴퓨터 게임에서도 Software로 각종 AI Algorithm을 구현하여 NPC(Non-Player Character)형태로 기존의 게임 Software에 내장되어 있다. 그런데 재미있는 것은 AI의 한 작은 부분인 Deep Neural Network이 최근의 지능혁명을 주도하고 있다는 사실이다. 90년대 초부터 인공지능의 실생활 응용이 우편번호 등의 자동 인식을 위한 OCR(Optical Character Reader)이나 냉장고와 세탁기를 저전력으로 제어하는 Fuzzy Logic 그리고 로봇 등에서 시작되고 있었다. 하지만 최근 들어 Data Mining 등의 이름으로 Google이나 Amazon, Apple 등과 같은 회사들이 Web에서 대량의 정보(Big Data)를 축적한 뒤 이를 인공지능 개념을 응용하여 상용화하는 연구를 진행함으로써 새롭게 인공지능 붐, 즉 지능혁명이 시작되었으며 이 때 가장 성공적인 인공지능 Algorithm이 Deep Neural Network이었다. 하지만 그 연산 처리가 복잡하여 기존의 CPU에서는 원하는 성능을 실시간으로 구현하는 것이 어려워 슈퍼컴퓨터나 GPU와 같은 특수 CPU가 필요하였다. 즉, AI, 시각 인식 등과 같은 연산들의 도입이 필요한데 이 때문에, AI 전용 Engine, 또는 AI Processor가 필요하게 된 것이다.

Ⅱ. 지능형 SoC

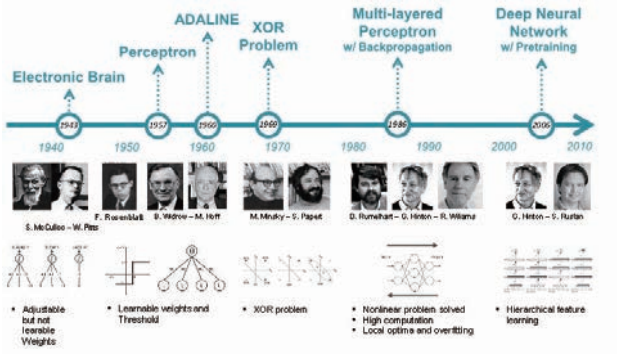
AI에서 다루는 지능은 환경변화에 적응하여 목적을 원만히 달성하는 능력이라고 이해하면 쉬울 것 같다. 이를 위해서는 주변 환경변화 감지, 목적달성 여부 판단, 이에 따른 행동 및 그 결과 분석, 결과 분석에 의한 자기 내부 수정 및 목적달성 방법 수정 등의 능력이 필요하다. 이 중에서도 가장 핵심이 되는 것이 Learning(학습)이며 기계(대부분은 컴퓨터를 가리킴)가 사람처럼 스스로 Learning하는 기술을 다루는 것이 바로 Machine Learning이다. Machine Learning은 Tom Mitchell의 정의에 따르면 "성능 지표(P)가 주어진 어떤 과업 (T)에 대한 경험(E)을 진행하면서 점차 P를 향상시키는 컴퓨터 프로그램"[1]이다. 이 분야는 80년대 후반부터 AI로부터 독립하여 하나의 연구 분야로 정착되어 SVM(Support Vector Machine) 등이 꾸준히 연구되어 오다가 최근 Deep Neural Network에 의해 전성기를 맞게 되었다.

지능형 반도체는 이러한 Machine Learning이나 AI 알고리듬을 소프트웨어만이 아니라 반도체 SoC로 구현하는 것을 말한다. 이것이 대두된 이유는 크게 1) 컴퓨터 구조의 한계 극복, 2) 핵심 문제가 "계산"에서 "인식"으로 변화, 3) 뇌연구의 진전 등을 들 수 있을 것이다. 먼저 최근 들어 CPU의 Clock 주파수 등 성능이 더 이상 개선되지 않고 있으며 Quad Core나 Octa Core 와 같이 한계에 다다른 CPU를 복수개 집적하는 방향으로 기술이 진보되

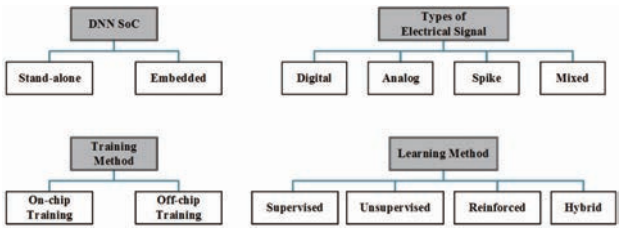
고 있다. 하지만 근본적으로 Von Neumann구조가 가지고 있는 메모리 Access에 대한 한계가 여전히 존재하고 있어서 저전력 고성능의 구현에는 제한적이다. 뇌를 모방하는 지능형 반도체를 이용하여 이를 극복하고자 하는 움직임이 크다. 두 번째로 Big Data의 대두와 함께 당면한 문제가 데이터의 계산 중심이던 종래의 정보시스템에 요구 기능이 점차 빅 데이터의 패턴 인식으로 바뀌고 있으며 이를 가속시키기 위한 반도체가 필요하게 되었다. 마지막으로 인간의 뇌에 대한 연구가 비약적으로 진보하여 뇌의 기작이 많이 알려지게 되었다. 1940년대에 현재의 컴퓨터 구조를 결정지었던 Von Neumann도 당시에 알려진 뇌의 기작을 구현하려고 시도하였으며 당시 사용할 수 있었던 기술들의 한계에 의해 현재의 컴퓨터 시스템으로 실현되었던 것이다. 그 후 뇌에 대해 축적된 지식과 경험을 바탕으로 비약적으로 발전한 반도체 기술과 합쳐서 새로운 컴퓨터 구조 또는 정보처리 시스템 구조를 고안할 필요가 있으며 그 실현 가능성도 높다.

본고에서는 Machine Learning 중에서 최근 가장 각광을 받고 있는 Deep Neural Network을 중심으로 그 SoC 구현에 대해 살펴보고 아울러 주요 AI 기법인 SVM 및 Fuzzy Logic용 반도체에 관련하여 살펴보기로 한다.

〈그림 2〉에서 보듯이 Neuron의 수학적 모델이 Pitts와 McCulloch 에 의해 제안된 이래 이를 전자 회로로 구현하려는 시도가 계속되었다. 하지만 진정한 IC 또는 SoC개념의 시작은 Caltech의 Carver Mead에 의해 다양한 Neural System들이 IC로 만들어 진 것이 그 시작이라 할 수 있으며 이는 그의 1989년 저서 Analog VLSI and Neural Systems에 잘 설명되어 있다. 1980년대 말부터 Memory를 바탕으로 Neural Network를 또는 Neural Network를 바탕으로 Memory를 구현하려는 시도, Analog회로로 구현하려는 시도, Digital 회로로 구현하려는 시도 및 Pulsed Mode (Spike 또는 Neuromorphic) 형태의 SoC구현 등 많은 연구들이 진행되었다. 이에 대한 연구결과는 최중호교수의 1995년 저서 "Neural Information Processing and VLSI"에 자세히 정리되어 있다. 하지만 Neural Network에 대한 연구 열기가 식어진 1990년대 말부터 2000년대 중반까지



〈그림 2〉 Deep Neural Network의 발전 과정



〈그림 3〉 지능형 SoC의 구분

Neural Network SoC에 대한 연구가 거의 이루어지지 않았으며 소수의 그룹에서 2000년대 중반부터 KAIST를 중심으로 다시 CNN[2-4], SVM[5] 및 Fuzzy[6-8] Logic에 대한 IC 구현들이 ISSCC에 발표되는 실정이었다. 또한 최근에 Deep Neural Network의 성공에 따라 2015년 이후부터 관련 논문이 대거 발표되는 등 다시 Neural Network SoC에 대한 연구가 활발하게 되었다.

국내에서는 KT 연구소의 한일송 박사팀에 의해 1991년 6월에 640개의 시냅스를 가진 지렁이와 거머리의 지능 수준의 뉴런칩을 개발한 것이 그 시작이었다. 이어 한 박사팀은 1993년 13만5천개의 시냅스를 가지는 파리 지능 정도의 '뉴런 칩'을 개발하였다. 하지만 이후 응용이 제한적이고 연구가 축소되는 세계적인 흐름에 따라 중단이 되었다. 이후 2008년 ISSCC에서 KAIST가 CNN을 내장한 '시각 인식 칩'을 발표[9]하면서부터 본격화되었고 AI 개념을 도입한 Augmented Reality SoC를 개발하여 이를 안경형 기기인 K-Glass로 구현하여 2014년 2월에 ISSCC에서 발표[10]하는 등 세계의 연구를 선도해 나아가고 있다. 특히 2017년 2월에 개최되는 ISSCC에서는 KAIST가 CNN-RNN을 모두 지원하며 사용자가 원하는 대로 임의의 Deep Neural Network를 구현할 수 있는

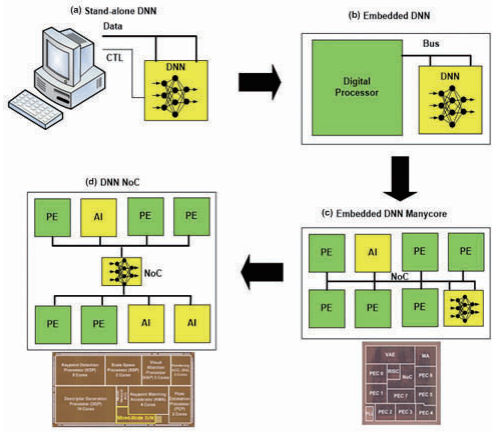
General Purpose Neural Processor Unit을 발표하며 또한 0.6mW라는 저전력에서 동작하는 CNN 칩과 이를 바탕으로 Wearable 얼굴인식기를 개발하여 발표하였다.

지능형 SoC는 전체 SoC에서의 지능 구현부(DNN회로)의 비율에 의해 〈그림 3〉과 같이 Stand-alone형과 Embedded형으로 나누어 생각해 볼 수가 있다. 이제까지의 대부분의 지능형 SoC는 Stand-alone형으로 시스템상에서 기존의 CPU 등과 함께 지능형 연산의 가속을 담당하는 형태를 가졌다. 하지만 KAIST에서는 Embedded DNN 또는 Embedded AI를 강조하여 전체 Many-core SoC내에서 일정 부분만을 DNN과 같은 AI Block이 차지하며 나머지 Block들은 Micro-processor와 같이 종래의 Von-Neumann 회로로 구현되는 구조를 주로 연구하고 있다. 또한 Digital회로로 구현하거나 Nonlinear기능의 구현이 손쉬운 Analog회로로 구현 또는 최근 Neuro-morphic형으로 관심을 끌고 잇는 Spike형태로도 구현이 가능하고 특히 Analog와 Digital을 혼용하여 사용하는 Mixed-mode형도 유망하다. 또한 SoC 상에서 사용중에 계속하여 Learning을 시키는 형태와 이미 외부에서 Learning된 Parameter들은 SoC에 이식하여 사용 중에는 학습을 못하며 이미 학습된 정보를 바탕으로 고성능 처리를 구현하는 Off-Chip Learning방식이 있다. 현재 대부분의 DNN회로들은 Off-Chip Learning형태이어서 학습은 불가한 Inference 기기로만 보아도 무방하다. 또한 학습 방식에 따라서도 외부에서 정답을 알려주면서 학습을 시키는 Supervised Learning, 외부에서 Data만을 주고 스스로 학습하여 그 Data들의 특징을 찾아 Category를 나누게 하는 Unsupervised Learning, 외부에서 정답을 알려주는 것이 아니라 회로가 추론한 결과에 대해서 옳은지 그른지와 같은 Reward만을 주는 Reinforcement Learning, 그리고 이들을 조합하여 사용하는 Hybrid방식이 있으며 현재의 대부분의 DNN에서는 Stand-alone, Digital형태의 Off-Chip Learning의 Supervised Learning방식이 사용된다. KAIST에서는 이뿐만이 아니라 Embedded, Mixed 형태의 On-Chip Learning 및 Reinforcement Learning 등 다양한 형태의 DNN 구현에 대한 연구도 계속하고 있다. 이제 다음 장에서는 몇 가지의 예를 들어 그 구체적인 구현방법에 대해 살펴본다.

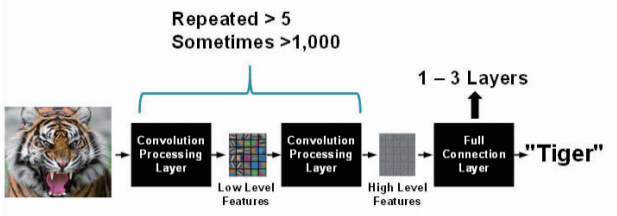
Ⅲ. 지능형 SoC의 구조와 구현 방법

1. Embedded AI 구조

Stand-alone형은 〈그림 4(a)〉에서 보듯이 기존의 컴퓨터에 Co-Processor형태의 AI가속기로서 설계된다. 이에 비해 Embedded AI는 SoC상의 다양한 IP Block들 중 하나로 AI가속기가 존재한다. 단순히 하나의 CPU와의 결합도 가능하지만 최근과 같이 Manycore Processor형에서는 〈그림 4(c)〉와 같이 하나 또는 몇 개의 Block들을 DNN 등의 AI 가속기로 대체한다. 또한 〈그림 4(d)〉와 같이 Manycore Processor를 연결하는 Network에도 Deep Neural Network이나 SVM 등의 AI Block들을 이용하여 Network Utilization이나 전체 IP Block들의 Load Balancing 등 지능형 제어를 할 수 있다. 〈그림 4(c)〉의 아래쪽 사진은 2008년 2월 ISSCC에서 발표되었던 KAIST의 지능형 SoC칩[11]으로 VAE라는 CNN을



〈그림 4〉 Embedded AI의 구조 변화



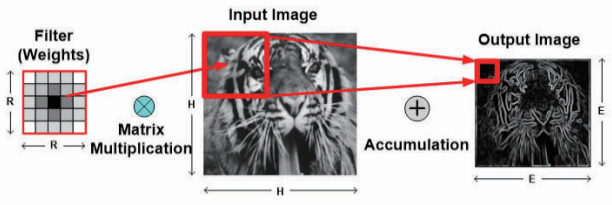
〈그림 5〉 CNN (Convolutional Neural Network) 의 이미지 인식 과정

기반으로 하는 지능 Block이 Multi-core Processor내에 집적되어 있는 구조이다. 〈그림 4(d)〉의 아래쪽 사진은 2014년 2월 ISSCC에서 발표되었던 KAIST의 지능형 SoC칩[12]으로 하단부의 SVM이 NoC를 Control하는 구조이다.

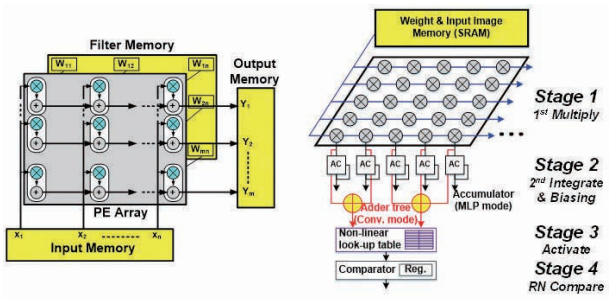
2. Digital CNN의 구현 예

최근의 Machine Learning의 대표는 Deep Neural Network이며 특히 CNN(Convolutional Neural Network)이다. 입력 Image를 받아서 Convolution 및 Pooling연산을 수~수백차례 반복하고 마지막으로 Full Connection Layer(예전의 Perceptron과 동일)들을 거쳐서 최종 인식된 출력을 내보내는 구조이다. 아이디어는 1980년대 말에 나왔지만 2010년에 종래의 다른 어떤 algorithm보다 Image인식에 매우 우수하다는 것이 밝혀진 이래 지금까지 인식이나 Machine Learning과 같은 AI Algorithm의 대표로 인식되고 있다.

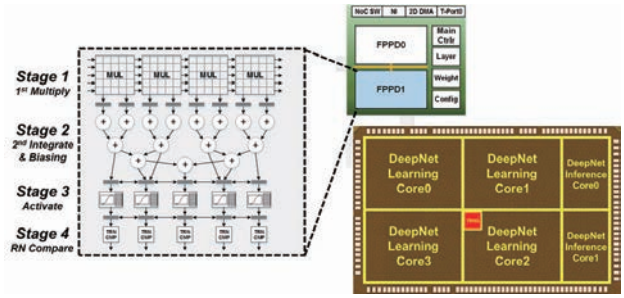
이 DNN은 Convolution연산이 반복되므로 Learning이 매우 복잡하여 수십 시간에서 1주일정도의 긴 학습기간이 필요하며 학습된 이후 입력 Image에 대한 인식출력을 얻어내는 Inference에도 GPU정도의 컴퓨터 성능이 필요하다. 이를 해결하고자 DNN용 가속기가 필요하게



〈그림 6〉 CNN의 Convolution 연산 수행 과정



〈그림 7〉 Convolution Layer와 Fully Connected Layer 모든 구현 가능한 설계



〈그림 8〉 세계최초의 저전력 CNN 칩 (KAIST)

되고 이에 대한 설계 기법을 여기에서 간단하게 소개하고자 한다.

CNN에서 Computing Power를 가장 많이 사용하는 연산이 Convolution이다. 이 Convolution 연산은 〈그림 6〉과 같이 R×R 크기의 Filter를 입력 Image의 일부분과 겹쳐서 Matrix Multiplication을 한 뒤 출력 Image의 일부분으로 표시한 뒤 Filter를 오른쪽으로 Sliding하여 다시 입력 Image의 일부분과 Matrix Multiplication을 하여 출력하는 동작을 반복하는 것이다.

이 연산에서 주로 사용되는 Matrix Multiplication은 주로 Image Pixel의 Row와 Filter의 Column Data간의 곱 및 이 곱셈결과들의 합으로 구성되어 있다. 결국 이러한 연산을 고속 및 저전력으로 구현하기 위해서는 Filter Size와 같은 R×R개의 Multiplier와 Accumulator가 필요하게 된다. 하지만 이렇게 많은 Multiplier + Accumulator (이제부터 PE라 부른다)를 하나의 칩 상에 집적하는 것이 어려우므로 PE의 개수를 줄이고도 연산 속도의 감소나 정확도의 감소를 줄여주는 것이 결국 Digital CNN 설계에서 중요하다. 〈그림 7〉은 KAIST가 2014년 ISSCC에서 발표한 실제적인 구현 예시이다. PE Array의 수를 줄이고 외부 Memory Access를 줄여 주어 작은 면적으로 고속이며 저전력으로 동작하도록 했다. 또한 5×5개의 16bit Multiplier가 구현되어 있으며 특별히 Accumulator부의 연결을 가변할 수 있도록 하여 CNN만이 아니라 Fully Connected Layer의 구현도 가능하도록 하였다.

〈그림 8〉은 2015년 ISSCC에서 KAIST가 발표한 세계최초의 저전력 CNN칩[13]을 보여준다. 칩에서 하나의 DeepNet Inference Core는 2개의 FPPD를 가지고 있

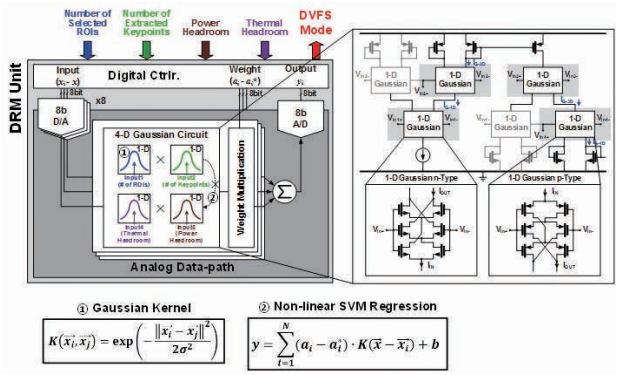
으며 하나의 FPPD 내에는 4개의 5x5 PE Array가 배치되어 있다. 따라서 하나의 DeepNet Inference Core에는 200개의 PE Array가 존재한다. 제작된 Chip은 185.3mW의 평균전력 소모와 411.3 GOPS의 성능을 보여서 GPU보다는 1.3배의 고속이며 일반 CPU보다는 42배의 고성능을 보였다.

3. Mixed Mode SVM의 구현 예

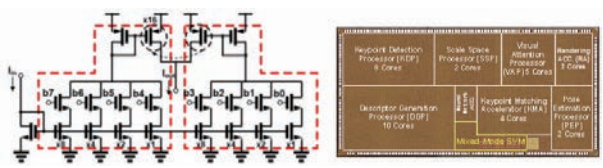
Support Vector Machine은 DNN의 고성능이 인정받기 전까지만 해도 가장 우수한 Classifier로 여겨졌으며 Big Data 처리가 필요치 않은 경우 현재에도 이용되고 있다. 기존의 Neural Network과는 달리 Kernel Method에 기초를 두고 Category별 Margin을 극대화시키는 Algorithm이다. Data를 Kernel Function, $K(x-x_i)$으로 변환하여 분류가 쉽도록 하는데 특히 Nonlinear Kernel (여기에서는 Gaussian Kernel)을 이용하면 매우 복잡한 분포를 갖는 Data들도 잘 분류할 수 있는 Nonlinear Support Vector Machine을 구현할 수가 있다.

〈그림 9〉에서 보듯이 Differential Amp형태의 Analog 회로를 이용하여 간단하게 Gaussian function을 구현하였다. 입력이 모두 4개이므로 4차원의 Gaussian Function이 필요한데 이를 위해 그림 오른쪽에서와 같이 4개의 Differential Amp를 cascade로 연결하여 이를 구현하였다. 상단의 Digital Controller에는 Weight Memory와 Learning Logic이 있어서 Learning에 의한 Weight Update가 가능하다. 따라서 Digital Learning 및 Analog처리, 즉 Mixed Mode Circuit으로 구현함으로써 전력 소모는 2.7배 감소할 수 있었고 면적은 1.6배 감소시킬 수가 있었다. 면적 감소가 1.6으로 작은 이유는 Digital Domain과 Analog Domain을 연결하기 위한 ADC/DAC가 추가되었기 때문이다. 하지만 Weight Multiplier의 경우 〈그림 10〉과 같은 Current Mode Multiplier를 사용하였기에 Digital 입력을 바로 받아서 Analog Multiplication이 가능하였다. 즉 8bit의 Weight가 b0~b7까지에 연결됨으로써 Iin이, 이 값에 비례하여 변형되어 출력 전류 Iout을 얻을 수가 있으므로 별도의 ADC와 Multiplier가 없이 하나의 간단한 회로로 구현이 가능하였다.

이렇게 Mixed Mode회로로 설계함에 따라 Digital회로만으로 설계된 Nonlinear SVM회로에 비해 면적이 1.6배 감소하였고 전력 효율은 2.7배 증가하였다.



〈그림 9〉 Mixed Mode SVM을 위한 Gaussian, SVM Regression 구현

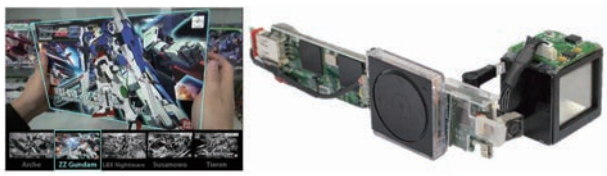


〈그림 10〉 Current Mode Multiplier를 사용한 Weight Multiplier

Ⅳ. 지능형 SoC의 응용 예

지능혁명이라고 불리는 만큼 지능형 SoC의 응용 영역은 매우 다양하다. 특히 최근의 Smart Machine, 자율자동차나 IoT 분야에서 응용이 커질 것으로 예상한다. 물론 스마트 안경이나 Watch와 같은 Wearable Device에서도 매우 활용도가 높을 것이며 가전제품이나 로봇 등에도 인공지능이 활용되어 사용성을 높일 것이다.

여기에서는 대표적인 응용으로 스마트 안경에 대해 살펴보기로 한다.



〈그림 11〉 KAIST에서 제작한 AR 안경, K-GLASS

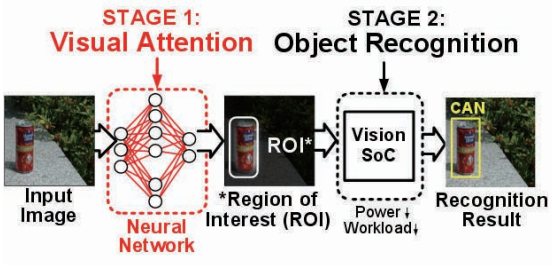
AR(Augmented Reality)가 최근 관심을 모으고 있다. AR은 실제 촬영된 화면과 Computer Graphics로 그린 그림이 하나의 화면에서 서로 상호작용을 하는 기술이다. 예를 들어, 〈그림 11〉에서와 같이 조립장난감의 박스를 들고 보기만해도 안경이 그 그림의 내용을 인식하고 조립된 장난감의 3차원 형상을 보여주어 그 장난감의 정보를 자세히 알 수 있도록 도와주는 것이 AR이다. 이러한 AR에서 필요한 연산은 물체의 인식 및 3차원 Graphics의 표시 등 종래의 스마트 폰에서 요구되는 화상처리 연산보다 월등히 복잡하다. 하지만 안경형태의 Form Factor때문에 battery의 크기는 훨씬 작아져서 구현이 어렵고 설사 구현이 되었다 하여도 배터리 사용 시간이 수십 분에 불과하여 매우 불편하다. 따라서 저전력, 고속으로 AR을 처리하기 위한 전용 Processor가 필요하며 이를 효율적으로 구현하는데 여기에서도 지능형 Block 들이 큰 역할을 담당하고 있다. 특히 Computing Workload감소와 Hardware Resource Utilization향상에 효과가 크다.

우선 Computing Workload를 줄여주는데 이용될 수가 있다. 〈그림 12〉에서와 같이 Pattern Recognition Pipeline에서 앞쪽 Stage에 CNN을 삽입하고 입력 Image의 Resolution을 낮추어 저전력으로 빠르게 Object Detection([14-19] 참고)을 행한다. 이 과정은 마치 우리 인간이 사진을 볼 때 모든 그림을 다 보는 것이 아니라 중요한 곳만을 주의 집중하여 바라보는 것과 유사하여 이 단계를 Visual Attention단계라고도 부른다. 이후 주요한 물체가 있는 영역 즉 ROI만을 좀 더 복잡한 CNN이나 Multicore Processor에서 고속으로 처리하여 물체를 인식할 수가 있다.

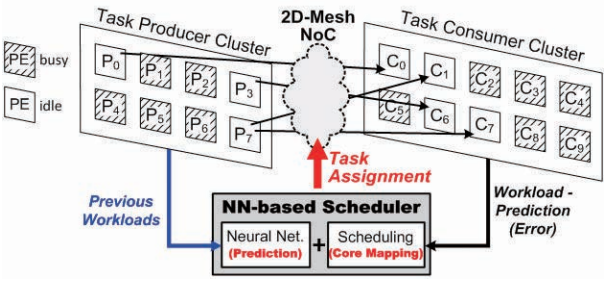
다음으로 Workload의 Balancing과 Hardware Utilization의 증가에도 인공지능이 효과적이다. 지능 Block이 이전 Workload에 대한 정보를 입력으로 받아 이후의 Workload상황을 예측하여 각각의 PE에 Workload를 재분배하며 Workload Monitor로 이를 Feedback받아 Learning을 통해 더욱 정확하게 Workload를 분배하는 것이 가능하다 (〈그림 13〉).

〈그림 14〉에서는 처리해야 하는 Keypoint 와 ROI를 정확히 예측하여 PE들을 할당하며, 또한 PE들의 Voltage와 Frequency를 변환하도록 하여 약 51%의 전력소모를 줄여 주는 것을 보여 주고 있다.

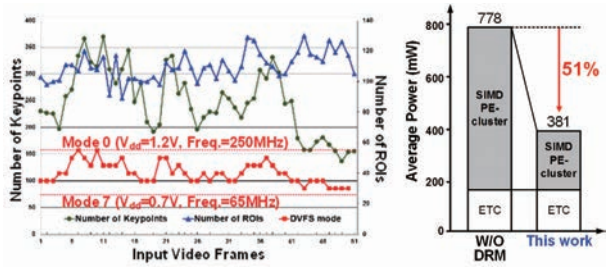
〈그림 15〉은 이 칩의 사진과 이를 스마트 안경용 보드에 조립된 모습을 보여 주고 있다. 제작된 칩은 1.22TOPS의 성능을 가지며 381mW의 평균전력 소모를 보이고 있다. 조립된 스마트 안경은 실내에서 책을 인식



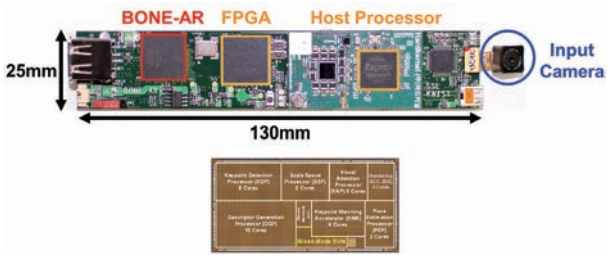
〈그림 12〉 KAIST에서 제작한 AR 안경, K-GLASS



〈그림 13〉 Workload Monitor의 Feedback을 통한 Workload 분배



〈그림 14〉 예측한 Keypoint ROI를 통한 PE 할당 결과



〈그림 15〉 K-GLASS의 칩과 조립된 보드

하여 정보를 제공해 주거나 책 속의 사진에 3차원 이미지를 표시해주어 사용자의 보는 각도에 따라 3차원 이미지의 방향을 바꾸어 주는 증강현실 기능을 잘 구현하였다.

Ⅴ. 전망과 결론

인공지능에 대한 관심이 증가하고 컴퓨팅에 대한 개념이 종래의 단순 Calculation에서 사물과 사건의 Recognition으로 변화되고 있다. 또한 인공지능에 대한 응용분야가 확대됨에 따라 소프트웨어만이 아니라 지능형 SoC도 본격적인 관심을 받기 시작하고 있다. 저전력 및 고속의 인공지능 응용에서는 소프트웨어보다는 반도체 칩으로 구현이 가격 및 성능이라는 측면에서 매우 우수하기 때문이다.

이러한 응용은 크게 Cloud Server용과 Mobile 단말용으로 나누어 볼 수가 있다. 이중 Cloud Server는 Stand-alone형이거나 Co-Processor형태의 지능형 SoC가 될 것이라 예상되며 이 분야에서는 MS, IBM 및 Google등의 여러 대기업에서 지능형 칩을 개발하고 있다. 특히 이들은 현재의 GPU를 대체하여 Data Server의 성능을 향상시키기 위해 범용성이 높은 지능 SoC를 개발하고 있다. 하지만 드론(Drone)과 자율주행자동차 그리고 로봇이나 IoT 단말 등과 같이 모바일 기기에서도 인공지능에 대한 요구가 높고 전력 소모 등에서의 제한 조건이 큰 만큼 이에 대한 지능 SoC에 대한 요구도 크다. 특히 이 분야에서는 범용성보다는 특정 응용에 특화된 Embedded 지능 SoC가 널리 이용될 것이기 때문에 이 분야를 집중적으로 연구한다면 한국이 강점을 가지고 있는 분야라고도 판단된다.

앞으로의 인공지능SoC에 대한 연구는 Cloud의 경우 인간이 풀지 못하는 여러 문제들을 풀어주는 Super-intelligence의 실현으로 전개될 것이며 Mobile의 경우 인간을 더욱 초능력자로 만들어 주는 즉, Super-man을 만들어 주기 위한 지능형 SoC로 발전하리라 예상된다. 이렇게 인간의 지적 능력을 도와주는 의미에서 AI가 Artificial Intelligence가 아니라 Augmented Intelligence의 뜻으로 해석될 수도 있다.

지능형 SoC분야의 연구와 개발은 개념이 잡히고 여러 연구자들이 앞 다투어 개발하기 시작하는 이제부터가 중요한 시점이라고 판단한다. CPU에서 DSP로 그리고 다시 GPU로 발전함에 따라 이제는 지능형 SoC, 즉, NPU(Neural Processing Unit)이 주류를 이룰 것이라 여겨지며 이 때 한국인이 한국에서 개발한 한국형 NPU가 세계에서 널리 사용되는 그런 계기가 되었으면 하는 바람이며 이에 대한 연구와 지원이 확대되기를 기대한다.

참 고 문 헌

[1] Tom M. Mitchell, Machine Learning, Mc Graw-Hill, p. 2, 1997, New York.

[2] Gyeonghoon Kim, et al. "A 1.22TOPS and 1.52mW/MHz Augmented Reality Multi-Core Processor with Neural Network NoC for HMD Applications", ISSCC 2014

[3] Seungjin Lee, et al. "The Brain Mimicking Visual Attention Engine: An 80x60 Digital Cellular Neural Network for Rapid Global Feature Extraction", SOVC 2008

[4] Jinwook Oh, et al. "An Area Efficient Shared Synapse Cellular Neural Network for Low Power Image Processing", VLSI-DAT 2009

[5] Youchang Kim, et al. "A 4.9mW Neural Network Task Scheduler for Congestion-minimized Network-on-Chip in Multi-core Systems", ASSCC 2014

[6] Junyoung Park, et al. "A 92mW Real-Time Traffic Sign Recognition System with Robust Light and Dark Adaptation", ASSCC 2011

[7] Minsu Kim, et al. "A 22.8GOPS 2.83mW Neuro-fuzzy Object Detection Engine for Fast Multi-object Recognition", SOVC 2009

[8] Seungjin Lee, et al. "A 92mW 76.8GOPS Vector Matching Processor with Parallel Huffman Decoder and Query Re-ordering Buffer for Real-time Object Recognition", ASSCC 2010

[9] Kwanho Kim, et al. "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual Attention Engine", ISSCC 2008

[10] Jinwook Oh, et al. "An Asynchronous Mixed-mode Neuro-Fuzzy Controller for Energy Efficient Machine Intelligence SoC", ASSCC 2011

[11] Donghyun Kim, et al. "81.6 GOPS Object Recognition Processor Based on a Memory-Centric NoC", TVLSI 2009

[12] Gyeonghoon Kim, et al. "A 1.22TOPS and 1.52mW/MHz Augmented Reality Multi-Core Processor with Neural Network NoC for HMD Applications", ISSCC 2014

[13] Seongwook Park, et al. "A 1.93TOPS/W Scalable Deep Learning/Inference Processor with Tetra-Parallel MIMD Architecture for Big Data Applications", ISSCC 2015

[14] Joo-Young Kim, et al. "A 66fps 38mW Nearest Neighbor Matching Processor with Hierarchical VQ Algorithm for Real-Time Object Recognition", ASSCC 2008

[15] Minsu Kim, et al. "A 54GOPS 51.8mW Analog-Digital Mixed Mode Neural Perception Engine for Fast Object Detection", CICC 2009

[16] Jinwook Oh, et al. "A 1.2mW On-Line Learning Mixed Mode Intelligent Inference Engine for Robust Object Recognition", SOVC 2010

[17] Junyoung Park, et al. "Online Reinforcement Learning NoC for Portable HD Object Recognition Processor", CICC 2012

[18] Injoon Hong, et al. "A 125,582 vector/s Throughput and 95.1% Accuracy ANN Searching Processor with Neuro-Fuzzy Vision Cache for Real-time Object Recognition", SOVC 2013

[19] Kyuho Lee, et al. "A Vocabulary Forest-based Object Matching Processor with 2.07M-vec/s Throughput and 13.3nJ/vector Energy in Full-HD Resolution", SOVC 2014

유회준

• 1983년 2월 서울대학교 공과대학 전자공학과 학사졸업
• 1985년 2월 한국과학기술원 전기및전자공학부 석사졸업
• 1988년 8월 한국과학기술원 전기및전자공학부 박사졸업
• 1988년 9월~1990년 12월 미국 Bell Communications Research 연구원
• 1991년 2월~1995년 2월 현대전자 반도체연구소 DRAM 설계실장
• 1995년 3월~1998년 1월 강원대학교 전자공학과 부교수
• 1998년 2월~현재 한국과학기술원 전기및전자공학부 교수
• 2001년 7월~2003년 6월 시스템반도체기반기술개발사업 설계기술 연구단장
• 2003년 9월~2005년 9월 정보통신부 IT SoC/차세대 PC PM
• 2007년 1월~현재 한국과학기술원 반도체시스템설계응용연구센터 소장
• 2009년 3월~2011년 3월 한국차세대컴퓨팅학회 회장
• 2008년 1월~현재 IEEE Fellow

〈관심분야〉
Intelligent Vision SoC, Deep Learning SoC, Wearable Healthcare, Bio-Medical SoC